Engagement Analysis Through Computer Vision

Zachary M MacHardy
Computer Science Department
University of North Carolina
at Chapel Hill
Chapel Hill, United States

Kenneth Syharath
Computer Science Department
University of North Carolina
at Chapel Hill
Chapel Hill, United States

Prasun Dewan
Computer Science Department
University of North Carolina
at Chapel Hill
Chapel Hill, United States

***Abstract*— As distributed online communication becomes increasingly common, and audiences for live online presentations grow larger, the ability to receive meaningful feedback from audience members who are distant and distributed becomes a necessity. To this end, we have built upon previous work to create a tool that is capable of providing real time feedback to an online presenter about the engagement level of the audience. The tool makes inferences by using computer vision and machine learning techniques to analyze the faces of audience members.**

Keywords*: *Computer Vision, Machine Learning, Distance Lecturing, Online Presentations

I. INTRODUCTION

As the world continues to advance into the digital age, and the demand for access to online education in particular and distributed presentations in general continues to grow, unabated [5], the need for a comprehensive set of tools for the online presenter grows apace. In both business and educational settings, the ability of a presenter to judge the effectiveness of their presentation and the attentiveness of their audience at a glance is paramount, and is notably lacking in modern tele-presentation technology. This paper focuses on just such a technology: a tool which allows presenters, at a glance, to gather information about their audience and adjust their presentations accordingly.

The use of live lecturing in distance education is somewhat less common than the use of recorded video; there is less overhead involved in delivering pre-recorded content and the use of recorded video allows students to approach material at their own pace. Often the choice between live, in person lecturing, and impersonal recorded lectures is considered a de-facto dichotomy. Yet live online lecturing, in providing an easy means of lecture recording and an opportunity for active lecturer-student interaction and adaptive lectures, allows for many of the advantages of both distance and on-campus education. But while offering the benefits of a traditional distance education course is easy, offering features that capitalize upon the unique capacity to provide a real-time mechanism for interaction between presenter and audience are largely lacking.

This is not to say that venues for the audience to synchronously communicate with the speaker have been ignored. For example, in the case of Fernando et al [10], Twitter allowed distance learners to use textual communication to ask questions in real time. But while the need for the audience to glean additional information from the speaker during a presentation might be assuaged by such technology, the reverse is not true. Such active communication can be a useful tool to a presenter, but it is difficult to parlay such irregular and active forms of communication into a more general concept of an audience's level of engagement. What can be accomplished by a glance around a room during a live lecture becomes seemingly impossible when one is presenting to a screen.

To address this problem, some modern online presentation tools, such as LiveMeeting, feature tools for audience members to indicate the level of their interest in a presentation. However, the use of this manual indication feature is rare.

By leveraging the ubiquity of webcam hardware in modern laptops and using an efficient facial analysis algorithm, this paper presents a solution that provides a distributed means of gathering the same information which might be the result of manual observation of audience members, in a way that is scalable to arbitrarily large audiences, and potentially allows speakers to adapt their lecturing even in the absence of active student communication.

This work is in the spirit of other research efforts that have also tried to infer user status from data captured about users and their activities. Fogarty et al [Fogarty] have shown that programmers' interaction with a software development environment can be analyzed to determine if they were interruptible. Carter and Dewan have shown that programmers' interaction can be analyzed to also predict if they are having difficulty. Similarly, research by Kapoor et al. [Kapoor] has shown that it is possible to reliably infer when kids, solving a Tower of Hanoi problem, are frustrated, by using cameras, posture seating chairs, pressure mice, and wireless Bluetooth skin conductance tests as sensors to collect data There has been substantial work on determining user emotions. The most recent and comprehensive research on this

COLLABORATECOM 2012, October 14-17, Pittsburgh, United States

DOI 10.4108/icst.collaboratecom.2012.250429

issue has been done by McDuff et al at MSR [McDuff]. They have developed techniques to determine three aspects of user emotion: valence (whether the emotion is positive or negative), arousal (degree of emotion), and engagement level. Their mining algorithms use data captured from several hardware sensors including microphone, camera, Kinect, wearable wrist sensor (sending electro dermal activity), and GPS. In addition, they used interaction data such as web URLs visited, documents opened, applications used, emails sent and received, and calendar appointments. They used the inferences mainly to allow users to reflect on their mood and recall events in the last week.

Our research is most closely to previous work by Benzaid and Dewan [12] which was the first to automatically determine the engagement level of a remote audience. It did so by analyzing videos of the audience members; specifically it used Viola-Jones [11] facial detection, light image manipulation, and support vector machine (SVM) to classify audience members into three states: Bored, Engaged, and Frustrated. A 5-person user study showed that this work was promising.

We have built on this work in three main ways. First, our user study includes a more diverse and large sample of individuals and targets scenarios closer to those that might occur in distributed lectures. **Second, our software**… **Thirdly our evaluation**… In the remainder of the paper we describe our work in more detail, answering the following questions: How reliable is the software in real-world application, and how practical is its use?

II. SOFTWARE

The means by which facial data is gathered and classified is relatively straightforward (though a more comprehensive examination can be read in the paper by Benzaid and Dewan [12]), and takes place in several major stages, which are outlined below:

1. Capture of facial information
2. Feature identification and image manipulation
3. (During set-up) Training of the SVM with pre-labeled data
4. (During classification) Classification by SVM of gathered information.

Facial data is gathered by means of an arbitrary webcam, in the case of this study, the onboard webcam of a laptop; Any camera capable of providing a basic two-dimensional video feed of users would be viable. During the training phase of the software, a recorded video of such data is reviewed by a user and manually tagged in places where they find themselves to be bored or engaged by the material they were observing.

Feature identification takes place using the Viola-Jones feature identification system, an implementation of which is available in the OpenCV library [2]. The identification works by using a series of classifiers called Haar cascades to search an integral image of the user's face for rectangular features in constant time. Benzaid and Dewan determined that three features most important to determining the level of engagement are the eyes, the mouth, and the nose.

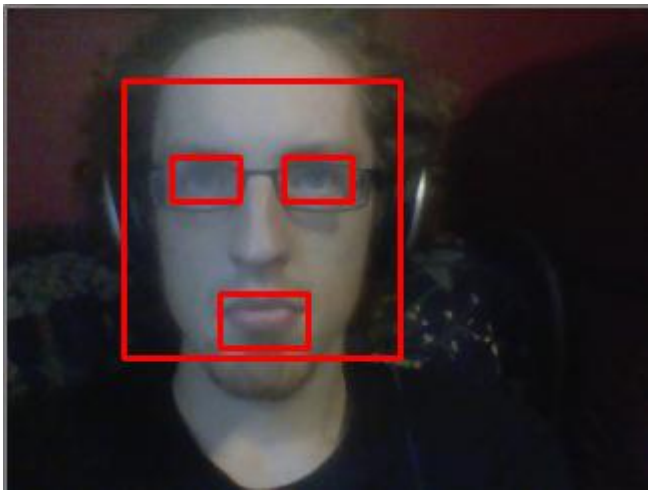

Figure 1: Detection of facial features.

In order to account for the variable size of images, due to factors such as the user's distance from the camera or the angle of the user's face relative to the camera, the images are resized via an affine transformation around a predetermined midpoint into standardized sizes on a 100x100 standard "affine face". This affine face image contains all important features at known locations, allowing for the extraction of precisely the information deemed important: the nose, eyes, and mouth. Each of these features is extracted from the affine face, grey-scaled and histogram-equalized, to allow for faster processing of data and remove possibly confounding information about skin tone and lighting. These grey-scaled images are then vectorized.

Figure 2: Affine face with detected affine features

During the training phase of the software, the program then proceeds to feed a series of these vectorized images along with a series of predetermined classifications, as either bored or engaged, into the training routine of a support vector machine (SVM). This machine works by constructing a hyperplane that separates points in each group placed in a high dimensional space in as distinct a way as is possible. During the classification phase of the software, new vectorized images

are projected into this high dimensional space, it is observed what side of the hyperplane they fall upon, and they are classified accordingly. The advantages of this method, while somewhat cumbersome during the training phase, lie in the efficiency with which points can be classified in real time after the SVM has been trained. We have additionally designed a system to perform such real time classification, yielding two-bit responses which encode a user's current engagement state.

The opposing classifications, Bored and Engaged, were chosen to represent the most extreme as well as the most useful metric that might be made available to a presenter about the state of an audience at a given time. Benzaid and Dewan included an additional frustrated state, but found that found it difficult to distinguish between boredom and frustration when tagging.

In order to further explore the usefulness of the technology, a user study of 20 diverse individuals was arranged. Thus this experiment was significantly larger than the 5 person user-study performed by Benzaid and Dewan. During the study, participants were instructed to watch about 30 minutes of video, taken from a series of TED talks [13] about a variety of subjects picked specifically to appeal to a spectrum of tastes broad enough that all viewers should ideally be bored and engaged at turns by subsets, though not identical subsets, of the videos. The choice of videos was a departure from the previous study by Benzaid and Dewan in which sitcom clips (with parts omitted to induce frustration) had been used to test the software. As the technology is intended for use in a presentation setting, we hypothesized that the changes in facial expression related to enjoying sitcom humor and those related to finding a lecture interesting were different enough to motivate a change of subject matter. Theoretically, the shift from interest to boredom would be somewhat harder to detect in a lecture setting. Additionally, it was predicted that users would not experience just one state during one lecture, but move from boredom to engagement and back again as each lecturer moved from topic to topic.

Specifically, the videos used covered topics ranging from art and music to biology and robotics;

III. RESULTS

Using a five-fold cross-validation test, on sets of data pre-labeled by participants, our results were somewhat mixed. Collection of data and, to a lesser extent, classification accuracy seemed to hinge largely on two factors:

- Physical appearance, in terms of accessories such as glasses, and the presence or lack of bangs obscuring the face
- The granularity with which users tagged their own moods

Taken in aggregate, the average accuracy of the classification was around 72%, which makes it as successful as respected research in both collaboration technology [Fogarty] and computer vision [Lana]. Most misclassifications were thanks to a tendency of the SVM to classify frames as engaged rather than bored, rather than vice versa. Though this is somewhat explained by the preponderance of frames toward engagement (the users were more often engaged than bored), it does not entirely account for this bias.

Participant 002		
Actual\Classified	Engaged	Bored
Engaged	0.96787	0.03212
Bored	0.10612	0.89387
Overall Accuracy:		0.93117
Participant 006		
Actual\Classified	Engaged	Bored
Engaged	0.96168	0.03831
Bored	0.77131	0.22868
Overall Accuracy:		0.71923

Figure 3: Example confusion matrices of subjects 002 and 006. Subject 002 possessed face-obscuring hair. Rows represent actual classifications, while columns represent predictions made by the software

Aggregate Data		
Actual\Classified	Engaged	Bored
Engaged	0.75265	0.24734
Bored	0.54297	0.45702
Overall Accuracy:		0.71941

Figure 4: Aggregate confusion matrix over all users. Bias toward classification of Bored frames as Engaged can be seen

Simply taking the results in aggregate, however, does not tell the entire story. Two confounding factors in particular, as listed above, affect the collection and classification of data.

These factors become more apparent as data is divided into general categories. As can be seen in Figure 5, the presence of hair in the face made a significant different in the machine's capability to classify moods. This is very likely because it is more difficult to observe small differences in facial features when those features are partially obscured, hamstringing the classification algorithm's ability to analyze with the data. The presence of eyeglasses, while ultimately having a much smaller effect on the classification accuracy, also contributes to this effect.

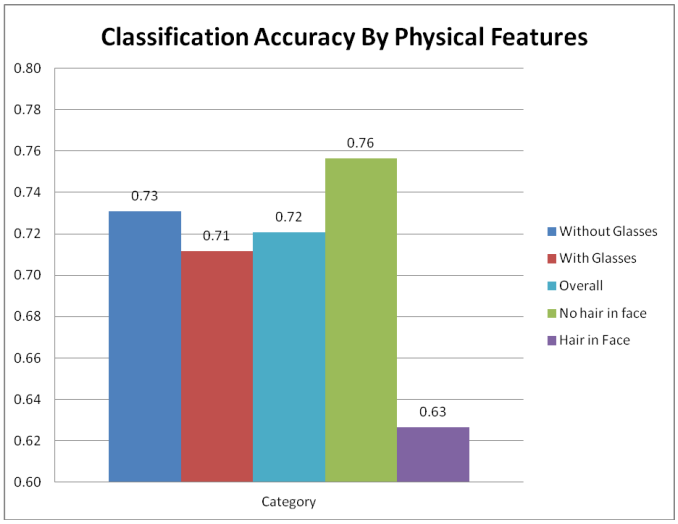

Figure 5: The classification accuracy (Ratio correct) as determined by physical features of the participant.

But, perhaps more important than the accuracy of the classification of recognizable images, is the effect such features have on the ability of the software to recognize the facial features present in an image at all.

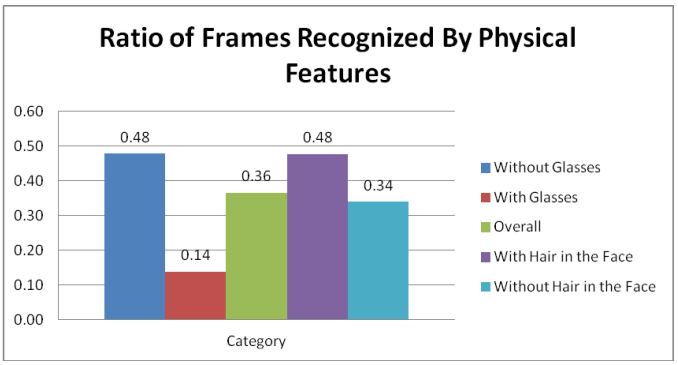

Figure 6: Ratio of frames containing machine-recognizable features to those without as determined by the physical features of the participant

As can be seen above, while eyeglasses contribute only a small effect to the classification accuracy, their presence greatly hampers the ability of the Viola-Jones algorithm to identify features in the first place. Interestingly, hair in the face seemed to have little to no effect on classification accuracy (It is worth noting that no participants had both hair in the face and wore eyeglasses, so the higher average of "With Hair in the Face" very closely mirrors "Without glasses")

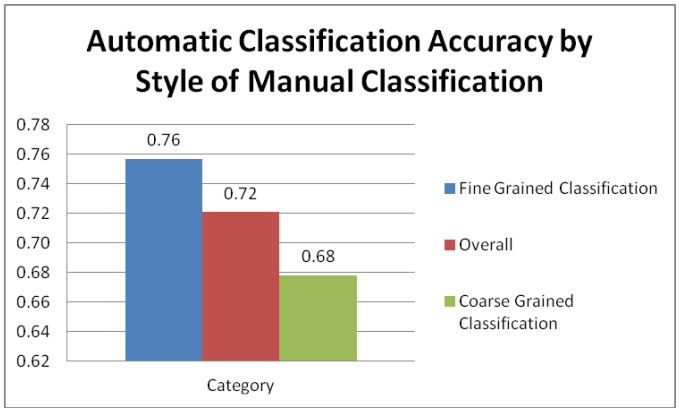

Figure 7: The classification accuracy of the software as determined by style of classification (Coarse v. Fine)

It is also the case that, unsurprisingly, the accuracy of the algorithmic classification was determined in part by the style of manual classification on the part of the study participant. For the purposes of this paper, a participant is identified as using coarse grained classification when found to have classified one or more (out of six) entire videos as either "bored or engaged", without allowing for neutral or opposing engagement states. While it is possible that a participant was genuinely engaged or bored throughout the entirety of the video, the correlation of such tagging with overall accuracy suggests that this predilection toward tagging in large chunks is worthy of note.

IV. Discussion

In performing this study we have identified the need for a system which provides passive feedback to online presenters, those in educational contexts specifically, but applicable to those in any context where distributed presentations are important. Having designed and tested such a system, we have demonstrated its potential to provide such real time feedback by utilizing an algorithm which leverages distributed client-side computation to determine and aggregate the engagement states of a large number of users in a scalable way. We have presented an exploration which has built upon previous work in the field and examined some possible concerns in distributing and implementing such technology, specifically with the accuracy of such a technology used in a more realistic setting, grappling with the real world constraints of confounding facial accessories and hairstyles. Though our total classification accuracy is lower than would be ideal, we assert that such a system, even based on current technology, would be of utility to online lecturer who desired passive audience feedback.

Some of the inaccuracies of the system are due to the physical features possessed by audience members, and may be, at least while still employing the Viola-Jones detection algorithm, difficult to address. Obscuration of the face, particularly by

eyewear, presents a challenge in feature recognition that is also difficult to overcome. In some cases, users who wore glasses were entirely missed by the facial detection algorithm: the software identified, out of around 8,000 possible frames, zero which contained a face. A more robust, but still efficient, algorithm for detecting facial features, perhaps less reliant on perfect detection, would likely go a long way to addressing this issue. The approach put forward by Halder et. al [1], using 36 facial action points to identify more general emotional states might be one such algorithm, if it is adapted to predict boredom and engagement, instead.

Another possible improvement might lay in way classifications are aggregated by a server. Currently, classifications are taken discretely, one at a time. When a frame is identified as bored, the client reports boredom. When interest is detected, the client reports interest. This can be problematic when the software erratically waffles between states. It is likely, instead, that users spend a reasonable portion of time being engaged with the material, then become bored, rather than waffling between states from second to second. A simple solution to this problem might be to record a sliding window of some arbitrary but reasonable number of classified frames, perhaps 25 or 50, and report, rather than the current frame's classification, the most common classification in the window. Such a change would smooth reporting and likely account for misclassification in the midst of a large block of one engagement level or another.

Tagging accuracy during the training phase of the software also had a significant impact on the performance of the software. This is an issue; if the machine is not given accurate data to begin with, it is impossible for it to construct an accurate SVM. It would be onerous to force the user to tag themselves with greater precision, and likely result in frustration on the part of the user. Rather, it might be possible, by aggregating a large number of users of diverse backgrounds, to construct a general model which would require less specific training per user to be effective.

Another improvement may come from an advance in hardware capability: As 3d cameras become less expensive to produce they have begun to replace traditional onboard 2d webcams in modern laptops. Leveraging the additional data made available by depth information may open the door to a much greater ability to recognize features despite partial facial obscuration, as well as allow for a more robust examination and analysis of important areas (eyes, nose, mouth) of a user's face [4].

Looking forward, in addition to algorithmic improvements, it would be constructive to test such software in the context of actual online lectures, in a classroom setting or otherwise. While the software has demonstrated the ability to provide passive information to an online presenter, it is unclear in what form that information would be best presented, or in what way such information would best be used. A study of lecturer use of such information might address such questions and open the door to the application of such technology in the virtual classrooms of the future.

[1] A. Halder et al, "Facial Action Point Based Emotion Recognition by Principle Component Analysis," Rorkee, Proceedings of the International Conference on Soft Computing for Problem Solving, 2011

[2] G. Bradski and A. Kaehler, "Learning OpenCV: Computer Vision with the OpenCV Library". O'Reilly Press, 2008

[3] G. Guo, S.Z. Li, K. Chan, "Face Recognition by Support Vector Machines," Grenoble, France: Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000

[4] G. Sandbach et al, "Static and Dynamic 3d Facial Expression Recognition: A Comprehensive Survey," London, UK: Image and Vision Computing, 2012

[5] I. E. Allen and J Seaman, "Going the Distance: Online Educ. in the United States" Wellesley, Massachusetts: Babson Survey Research Group, 2011

[6] M. Castrillón, et al, "Real-time detection of multiple faces at different resolutions in video stream". Las Palmas de Gran Canaria, Spain: 2007

[7] M. Ebner, "Introducing Live Microblogging: How Single Presentations can be Enhanced by the Mass," San Diego, California: Journal of Research in Innovative Teaching, 2009

[8] M. Yeasin et al, "Recognition of Facial Expressions and Measurement of Levels of Interest From Video," IEEE Transactions on Multimedia, vol. 8, no. 3, 2006

[9] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines," Cambridge, UK: Cambridge University Press, 2000.

[10] N.J.S. Fernando et al, "Live Lecture Streaming for Distributed Learning," London, UK: Scanning the Horizons: Institutional Research in a Borderless World Higher Education Institutional Research Network Conference, 2011

[11] P. Viola and M. Jones, "Robust Real-Time Face Detection" International Journal of Computer Vision, 2004

[12] S. Benzaid and P. Dewan, "Semantic Awareness through Comput. Vision" Chapel Hill, North Carolina: Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems, 2010

[13] Technology, Entertainment, Design Ideas Worth Spreading. www.ted.com

[14] Fogarty, J., et al., "Predicting human interruptibility with sensor" ACM Trans. Comput.-Hum. Interact. , 2005 **12** (1): p. 119-146

[15] Carter, J. and P. Dewan. "Design, Implementation, and Evaluation of an Approach for Determining When Programmers are Having Difficulty," Proc. Group 2010.

[16] Kapoor, A., Burleson, et al., "Automatic Prediction of Frustration," International Journal of Human-Computer Studies, 2007. 65(8).

[17] McDuff, D., et al. "AffectAura: An Intelligent System for Emotional Memory," in Proc. CHI. 2012.